BIOMETRICS 58, 163–170
March 2002

Two-Stage Designs for Gene–Disease Association Studies

Jaya M. Satagopan,[1,*] David A. Verbel,[1] E. S. Venkatraman,[1] Kenneth E. Offit,[2] and Colin B. Begg[1]

[1]Department of Epidemiology and Biostatistics, [2]Clinical Genetics Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, U.S.A.
[*]*email:* satago@biosta.mskcc.org

SUMMARY. The goal of this article is to describe a two-stage design that maximizes the power to detect gene–disease associations when the principal design constraint is the total cost, represented by the total number of gene evaluations rather than the total number of individuals. In the first stage, all genes of interest are evaluated on a subset of individuals. The most promising genes are then evaluated on additional subjects in the second stage. This will eliminate wastage of resources on genes unlikely to be associated with disease based on the results of the first stage. We consider the case where the genes are correlated and the case where the genes are independent. Using simulation results, it is shown that, as a general guideline when the genes are independent or when the correlation is small, utilizing 75% of the resources in stage 1 to screen all the markers and evaluating the most promising 10% of the markers with the remaining resources provides near-optimal power for a broad range of parametric configurations. This translates to screening all the markers on approximately one quarter of the required sample size in stage 1.

with breast cancer risk. To address these questions, investigators have planned a study to compare the frequencies of polymorphisms among breast cancer patients with and without an inherited *BRCA* mutation. The study subjects will be women affected with breast cancer. The association between a BRCA mutation and an SNP will be determined by recording the presence of the mutation and the polymorphism in these study subjects. This test will be carried out for many SNPs. Approximately 1500 anonymous SNPs and several known candidate polymorphisms will be analyzed and their frequencies compared among the two groups of individuals (patients with and without a *BRCA* mutation). The polymorphic variants to

jects will not be feasible. In this case, the one-stage design would involve evaluating all the m markers on T/m individuals. However, this can be inefficient in resource utilization since it may require large numbers of evaluations of genes that can be identified early in the study as extremely unlikely to be the true disease gene.

Consider, instead, optimization of the following two-stage design. In stage 1, screen all m genes on a set of n_1 individuals using the test statistic where the numbers of cases and controls in this subset of n_1 are chosen in proportion to their relative frequency in the full set of available subjects. Rank the genes based on the absolute value of the test statistic. Select the top ith proportion of these genes, i.e., select the

The probabilities P_1 and P_2 (equations (6) and (7)) can be evaluated using Monte Carlo simulation for given values of i, j, and μ.

6. Power Function for Correlated Gene Outcomes

In practice, the assumption of independent gene outcomes within subjects may not be even approximately true when testing multiple markers. Gene outcomes can be correlated due to various phenomena such as genetic linkage and loss of heterozygosity (evolutionary causes) and allele frequency and marker density (recombination). Correlation (denoted ρ) due to recombination can be easily quantified (Feller, 1966). Here we focus only on the aggregate correlation rather than correlation due to specific causes.

Under the assumption of independence, the true gene outcomes have a mean of μ, while the null gene outcomes have a mean of zero. However, when we cannot assume independence, the null genes in the neighborhood of the true genetic locus need not have a mean of zero since the mean outcome will be influenced by the correlation between the null and the true genes. Therefore, the mean outcome of the null genes will reduce to zero as a function of correlation as one moves away from the neighborhood of the true gene.

As defined in the previous section, let (X_1, X_2) denote the true gene outcomes in stages 1 and 2, respectively, normally distributed with mean $(\mu n_1, \mu n)$ and covariance matrix Σ (given by equation (4)). Further, let $Y_{1,u}$, $u = 1, \ldots, m-1$, denote the linear ordering on the genome of the null gene outcomes under stage 1. Similarly, let $Y_{2,u}$, $u = 1, \ldots, mi-1$, denote the outcomes of the selected null genes in their linear order to be evaluated in stage 2. The true gene (having outcomes X_1 and X_2 in stages 1 and 2, respectively) can be located anywhere along the genome. When addressing the design question, we consider the simple case where we assume that the correlations between adjacent pairs of loci are equal. As stated earlier, the true gene has mean μn_1 and variance n_1 in stage 1 and mean μn and variance n after stage 2. Therefore, the uth null gene away from the true gene has mean $\mu n_1 \rho^u$ and variance n_1 in stage 1 and mean $\mu n \rho^u$ and variance n after stage 2. The mean of each of the null genes approaches zero as the correlation between the true and the null genes decreases.

The power $P = P_1 \times P_2$ for this setting can be described as follows. In stage 1, P_1 is the probability that X_1 is among the top mi gene outcomes. Let $g^*(\cdot)$ denote the density of $Y_{(m-mi)}$, which denotes the $(m-mi)$th ordered null gene outcome in stage 1. The density $g^*(\cdot)$ depends on the mean μ, sample size n_1, and the correlation ρ.

Therefore, the probability P_1 can be written as

$$P_1 = \int_{-\infty}^{\infty} g^*(y)[1 - F_1(y; \mu n_1, n_1)]dy. \tag{8}$$

P_2 is the probability that X_2 is greater than each of the $m - mi$ null gene outcomes in stage 2, conditional upon the results of stage 1. Hence, P_2 can be written as

$$P_2 = P\left(X_2 > \max_{1 \le u \le m-mi} \{Y_{2,u}\} \mid X_1 > Y_{(m-mi)}, \min_{1 \le u \le m-mi} \{Y_{1,u}\} > Y_{(m-mi)}\right). \tag{9}$$

As in the previous case, the probabilities P_1 and P_2 can be evaluated using a Monte Carlo simulation for varying values of i, j, and μ.

7. Results

7.1 *Optimal Two-Stage Design for a Single True Gene*

The power function discussed in the previous section can be used to provide guidelines for optimizing the study design. The power function can be maximized with respect to i, the proportion of genes selected for validation, and j, the proportion of resources allocated for stage 1, for given values of T, m, and μ. Further clarification of resource allocation is possible by expressing it in the context of the total sample size and the proportion of individuals allocated to stage 1. The number of individuals in a one-stage design is given by T/m, and that of a two-stage design is $[j + (1-j)/i]T/m$. The ratio of the number of individuals required for a two-stage design to the number required for the one-stage design, for fixed T and m, is thus given by $j + (1-j)/i$. Note that the proportion of individuals in a two-stage design allocated to stage 1 is given by $ij/(ij + 1 - j)$.

In our simulations, power is calculated for $\rho = 0, 0.10, 0.20, 0.40, 0.60, 0.80, 0.90$, and 0.98, where ρ is the correlation between adjacent genes and $\rho = 0$ corresponds to the case of independent gene outcomes. For the purpose of our simulations, the signal μ is calculated for cases where a one-stage design testing independent markers will have 30, 40, 50, or 60% power. Table 1 summarizes the results of these simulations for $m = 3000$ and $T/m = 5000$. Row (a) gives the maximum power of the two-stage design. The numbers in parentheses in row (b) give the design parameters i and j at which the maximum power is obtained. Figures 1 and 2 provide a graphical representation of $m = 1000$, $T/m = 500$, $\mu = 0.120$, and $m = 100$, $T/m = 100$, $\mu = 0.275$, respectively. The bold line in the figures give the maximum power of the two-stage design. The design parameters i and j at which the maximum power is obtained are shown below the horizontal axis. As correlation between the genes decreases, the power of the optimal two-stage design tends toward that of the independent gene outcomes for all combinations of T, m, and ρ. Further, for fixed correlation, power increases as the signal (μ) increases.

The results show that over a broad range of values of T, m, and μ in the case of independent gene outcomes ($\rho = 0.0$), the optimal design parameters are in the range of $i \in (9\%, 15\%)$ and $j \in (63\%, 76\%)$. The power of this optimal design is very close to a design where $i = 10\%$ and $j = 75\%$. Therefore, as a general rule, when the genes are independent, sufficient power can be obtained by allocating approximately 75% (j) of the resources for screening in stage 1 and by validating the top 10% of the genes (i) in stage 2. In the case of correlated gene outcomes, the optimal design parameters are in the range of $i \in (1\%, 22\%)$ and $j \in (52\%, 82\%)$. Applying the above rule-of-thumb design to the correlated gene outcome case, we find that this design provides a sufficient approximation to the

Table 1

Power of one- and two-stage designs for $m = 3000$, $T/m = 5000$, *and values of* $\mu = 0.120, 0.130, 0.145,$ *and* 0.155 *for increasing values of correlation between adjacent markers. Row (a) gives the maximum power of the optimal two-stage design. Row (b) gives the optimal parameters* (i, j). *Row (c) gives the power corresponding to a rule-of-thumb two-stage design (when* $i = 0.10$ *and* $j = 0.75$). *Row (d) gives the power of a one-stage design. Row (e) gives the power (and percentage increase in cost) when using a one-stage design where the total number of individuals is fixed.*

	Correlation (ρ)

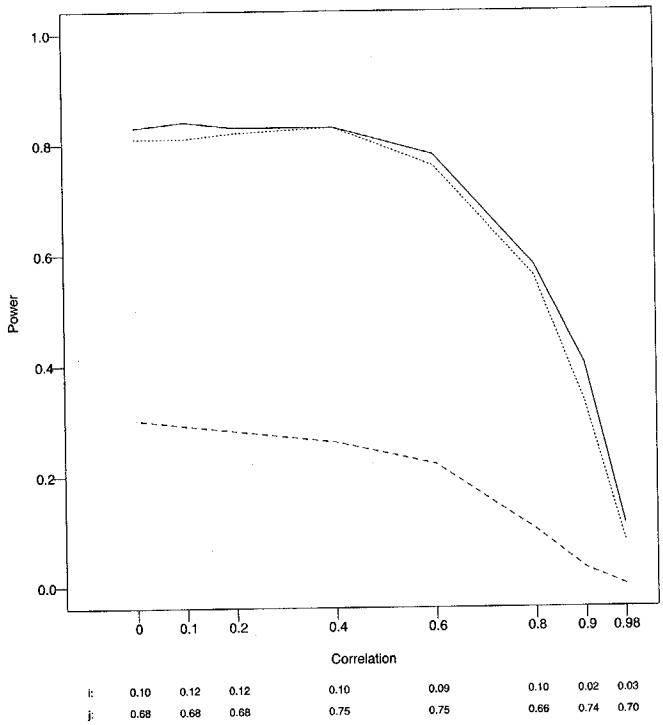

Figure 1. Power of one- and two-stage designs for $m = 1000$, $T/m = 500$, and values of $\mu = 0.120$ for increasing values of correlation between adjacent markers. The bold line shows the maximum power of the optimal two-stage design. Optimal parameters (i and j) are shown below the horizontal axis. The dotted line shows the rule-of-thumb two-stage design (when $i = 0.10$ and $j = 0.75$). The dashed line gives the power of a one-stage design. The value of $\mu = 0.120$ corresponds to a

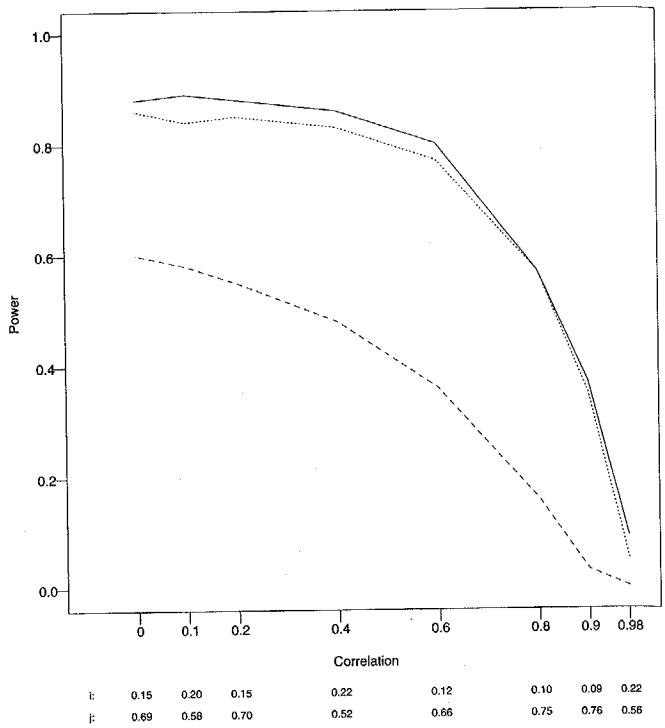

Figure 2. Power of one- and two-stage designs for $m = 100$, $T/m = 100$, and values of $\mu = 0.275$ for increasing values of correlation between adjacent markers. The bold line shows the maximum power of the optimal two-stage design. Optimal parameters (i and j) are shown below the horizontal axis. The dotted line shows the rule-of-thumb two-stage design (when $i = 0.10$ and $j = 0.75$). The dashed line gives the power of a one-stage design. The value of $\mu = 0.275$ corresponds to a

Table 2

Power to detect all five true genes of association using one- and two-stage designs in the presence of $m = 3000$, 1000, and 100 independent markers and $T/m = 5000$. Column (a) gives the maximum power of the optimal two-stage design. Column (b) gives the optimal parameters (i, j). Column (c) gives the power corresponding to a rule-of-thumb two-stage design (when $i = 0.10$ and $j = 0.75$). Column (d) gives the power of a one-stage design.

m	μ	(a)	(b)	(c)	(d)
3000	0.061	0.98	(0.09, 0.82)	0.96	0.30
	0.064	0.99	(0.12, 0.80)	0.98	0.40
	0.066	0.99	(0.12, 0.81)	0.98	0.50
	0.069	0.99	(0.15, 0.75)	0.99	0.60
1000	0.056	0.95	(0.12, 0.82)	0.92	0.30
	0.059	0.98	(0.12, 0.80)	0.95	0.40
	0.062	0.98	(0.19, 0.76)	0.97	0.50
	0.065	0.99	(0.12, 0.85)	0.98	0.60
100	0.046	0.76	(0.14, 0.81)	0.56	0.30
	0.0485	0.81	(0.14, 0.85)	0.67	0.40
	0.051	0.87	(0.14, 0.85)	0.73	0.50
	0.054	0.91	(0.14, 0.85)	0.80	0.60

baseline cost per chip), then the total cost of the study given by equation (1) would be modified as $T = n_1 m + n_2 mi + C \times (n_1 + n_2)$. Maximizing the power using this cost function could alter the optimal design parameters. However, note that $T = n_1 m(1+C/m)+n_2 m(i+C/m)$. The fraction C/m represents the relative cost of ascertaining an individual to the cost of genotyping that individual (m, the total number of markers evaluated per study subject, is the total cost of genotyping an individual, assuming a unit cost for each marker genotype). If $C \ll m$, then $T \approx n_1 m + n_2 mi$, and the results presented in the previous section can be applied. Another issue contributing to C could be the availability of sufficient cases, particularly when the disease is rare.

If the total number of individuals (N) is fixed (and m, the total number of genes, is given), then the optimal design is to perform all m gene studies on every individual. It is pertinent to pose the following question in this setting. How much power do we lose by using our rule-of-thumb design versus performing all m gene studies on all N individuals? This is

expect the test statistics of 3000 equally spaced markers to be more correlated than 100 equally spaced markers in a fixed genomic region. For studies of isolated populations where linkage disequilibrium extends across a distance of 30–50 kilobases (i.e., correlation between loci in a distance of 30–50 kilobases), it can be anticipated that less than 100,000 markers will be required to identify candidate regions of gene/disease association (Boehnke, 2000). Therefore, having 3000 equally spaced markers over the entire genome would result in markers with very low correlation. While the actual correlations can only be estimated from the observed data at the end of the study, broader assumptions about the correlations must be used in the setting of study designs. Often these assumptions can be based on *a priori* knowledge about the markers from previous studies, if such information is available.

After examining Table 1 and Figures 1 and 2, the similarity in power between the optimal two-stage design and the rule-of-thumb design is clearly shown. Furthermore, it is clear that the one-stage design has much lower power. Therefore, when the principal design constraint is total cost, as represented by the total number of gene evaluations, the rule-of-thumb two-stage design gives a pragmatic approach that provides most of the power achieved by a one-stage design at a fraction of the cost.

Acknowledgements

The authors would like to thank two referees and an associate editor for their insightful comments. This research was supported in part by National Institutes of Health grants R01 GM60457 and CA73848.

Résumé

Le but de cet article est de décrire une stratégie d'étude à deux étapes qui maximise la puissance de détection d'associations gène-maladie quand la principale contrainte est le coût total, représenté par le nombre total d'évaluations de gènes plutôt que le nombre total d'individus. Dans la première étape, tous les gènes d'intérêt sont évalués sur un sous-groupe d'individus. Les gènes les plus prometteurs sont alors évalués sur d'autres sujets dans la deuxième étape. Ceci évitera de gaspiller du matériel sur des gènes ayant une faible probabilité d'être associés à la maladie d'après les résultats de la première étape

BRCA2 in breast cancer families. The Breast Cancer Linkage Consortium. *American Journal of Human Genetics* **62,** 676–689.

Martin, E. R., Kaplan, N. L., and Weir, B. W. (1997). Tests for linkage and association in nuclear families. *American Journal of Human Genetics* **61,** 439–448.

Satagopan, J. M., Offit, K., Foulkes, W., Robson, M. E., Wacholder, S., Eng, C. M., Karp, S. E., and Begg, C. B. (2001). The lifetime risks of breast cancer in Ashkenazi Jewish carriers of BRCA1 and BRCA2 mutations. *Cancer Epidemiology, Biomarkers, and Prevention* **10,** 467–473.

Schaid, D. J. (1996). General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology* **13,** 423–449.

Schaid, D. J. and Rowland, C. (1998). Use of parents, sibs, and unrelated controls for detection of association between genetic markers and disease. *American Journal of Human Genetics* **63,** 1492–1506.

Teng, J. and Risch, N. (1999). The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Research* **9,** 234–241.

Received November 2000. *Revised October* 2001. *Accepted October* 2001.